La datavisualisation comme outil pour la recherche académique

Datavisualization as a tool for academic research

Clara Galliano – David Reymond – Luc Quoniam

English Abstract – Web development and digitization have allowed massive access to information and data. Today, many tools available make it possible to highlight the relationships between disciplines, between research structures and between the researchers themselves. Mapping is the fundamental tool for the development of information system to represent data in visual form (graph, network, map). We speak of "datavisualization". This poster is part of the research work of a PhD thesis in information and communication sciences.

1 INTRODUCTION

L'essor d'internet et du numérique ont permis la démocratisation de l'information grâce à la disponibilité croissante des données rendue possible par le développement des "Nouvelles technologies de l'information et de la communication" (NTIC).
Ces avancées technologiques ont fait émerger des mouvements comme celui de la "science ouverte" (traduit de l'anglais "Open Science"), dont en particulier certaines composantes de celle-ci autour des "données ouvertes" (Open Data) et de "l'accès libre" (Open Access).

Ces mouvements questionnent donc les sciences au sens large, et notamment les sciences humaines et sociales (SHS), sur de nouveaux objets de recherche et de nouvelles méthodes (datamining, traitement automatique des langues, text mining, cartographie…).
Le fait de cartographier les connaissances scientifiques n'est pas récent. L'organisation des connaissances puise son origine de l'Encyclopédie de Diderot et d'Alembert (éditée de 1751 à 1772). Depuis, de nombreuses méthodes d'organisation, de classification [1], d'indexation et de représentation ont vu le jour (classification de Dewey, carte de Bernal, carte d'Ellingham…).
De ces travaux est née la scientométrie, sur une approche cette fois-ci quantitative liée aux bases de données documentaires (création d'index de citation qui sera utilisé plus tardivement pour l'évaluation de la recherche et des chercheurs). Mesurer la science revient donc à mesurer l'activité scientifique [2].

2 ETUDE DE L'EXISTANT

Aujourd'hui, les techniques cartographiques peuvent s'appliquer à toutes les disciplines scientifiques. De nombreuses études ont été réalisées, dont notamment en médecine pour les sciences dites "dures" ou "exactes" [3], et pour les sciences "molles" (sciences humaines) [4].
Dans notre cas, nous resterons dans la discipline scientifique des Sciences de l'Information et de la Communication (SIC).
Un travail récent dirigé par Luc Quoniam sur la cartographie de la recherche académique au Brésil (figure 1) nous a inspiré cette nouvelle étude. Nous avons souhaité appliquer une partie de sa recherche pour le cas de la France.

3 METHODOLOGIE

Nous nous sommes basés sur les travaux antérieurs déjà réalisés dans notre discipline – les SIC- et d'autres domaines scientifiques.
Comme dit précedemment, la cartographie et la visualisation des données sous forme de réseaux ont déjà été envisagées dans plusieurs disciplines (autres que les SHS).

Dans notre cas, en plus de mettre en réseau les structures de recherche et les chercheurs, nous avons pensé compléter notre démarche de valorisation scientifique et de circulation des connaissances, en ajoutant d'autres critières :

- Production scientifique (communication, article, ouvrage, chapitre…)
- Lien entre chercheurs et leurs pairs
- Réseau de citation (à charge ou à décharge de l'auteur)
- Recherche par mots-clés, par discipline
- Lien entre structures, institutions
- Classement interdisciplinaire des thèses (sur la base des thèses françaises) à partir de la CIB.

Un travail sur la collecte et le traitement des données est fait en amont de la cartographie grâce au langage de programmation Python.

Cette étape nous permet de nettoyer les données comportant des erreurs, des coquilles et les doublons en fonction de la construction de notre corpus.
Pour tous les critères évoqués précédemment, les langages de programmation JSON, JavaScript (D3.js), HTML et CSS sont également nécessaires pour la réalisation finale et l'affichage des résultats.
Nous avons ensuite utilisé le logiciel libre de visualisation Gephi pour cartographier et développer notre réseau.
La méthodologie de recherche s'inspire entre autre de celle qui a été établie pour le cas du Brésil (cartographie réalisée à partir de la plateforme "Lattes" pour obtenir la production scientifique des unites de recherche).
En ce qui concerne le sujet des thèses françaises, nous avons choisi la base nationale des thèses françaises (theses.fr) et la base européenne des documents brevets (OEB) qui se sert de la classification internationale des brevets (CIB).

4 Resultats

Les points détaillés ultérieurement sont en cours de réalisation et font l'objet d'un travail de recherche dans le cadre d'une thèse.

Le dernier point concernant l'indexation des connaissances pour faire dialoguer les disciplines dans le principe de l'interdisciplinarité est validé.
Notre expérimentation s'est appliquée ici sur le domaine de recherche « eau ».
A partir de la précision du classement hiérarchique du schéma de la CIB [5], nous obtenons le résultat suivant :
eau > discipline > classe (CIB) > sous-classe > titre de thèse

Grâce à l'indexation des résumés de thèse permise par la classification de la CIB à partir d'une requête précise, nous avons proposé les visualisations suivantes :

- Cercle et cercle zoomable
- Nested Treemap (treemap imbriqué)
- Sunburst (figure 2)
- TidyTree.

5 Conclusion

Le but de ce travail de recherche sera dans un premier temps de dresser une cartographie complète de la discipline des Sciences de l'Information et de la Communication à partir de plusieurs indicateurs bibliométriques.
Cette visualisation permettra de mettre en evidence les relations scientifiques au sein d'une même discipline, de développer un véritable outil d'aide à la création d'un l'état de l'art dans un champ établi. Vous l'aurez compris, ce premier jet se concentre davantage sur les SIC, mais tend à se developer pour les autres champs disciplinaires de la recherche scientifique et académique.
Dans cette perspective, l'objectif sera de rendre cette recherche interopérable afin de pouvoir réutiliser la méthodologie et les résultats pour d'autres disciplines.

De plus, l'utilisation de la CIB comme pivot à la recherche académique ou l'indexation de connaissance permet de montrer le lien entre l'information académique et l'information technique et industrielle (brevet).
Les limites de cette experimentation sont telles :

- Homogénéisation des données
- Différences de citations selon les auteurs
- Problèmes d'affiliation des auteurs
- Classification des disciplines.

6 References

[1] Hudon, M. & El Hadi, W. (2010). Organisation des connaissances et des ressources documentaires : De l'organisation hiérarchique centralisée à l'organisation sociale distribuée. *Les Cahiers du numérique*, vol. 6(3), 9-38. https://www.cairn.info/revue-les-cahiers-du-numerique-2010-3-page-9.htm.

[2] Jeannin, P., Mouton, M-D. (2003). Vers une cartographie de la recherche en Sciences humaines et sociales : l'exemple de l'Ethnologie-Anthropologie sociale et culturelle. Politiques et management public, vol. 21(3), 101-120

[3] Lrhoul, H., Chartron, G., Bachr, A. & Benammar, O. (2015). La datavisualisation comme outil de pilotage de la recherche scientifique médicale au sein de la Faculté de Médecine et de Pharmacie de Casablanca. In : Évelyne Broudoux éd., *Big Data - Open Data : Quelles valeurs ? Quels enjeux : Actes du colloque « Document numérique et société », Rabat, 2015* (pp. 165-181). Louvain-la-Neuve, Belgique: De Boeck Supérieur. doi:10.3917/dbu.chron.2015.01.0165.

[4] Gallot, S. (2014) « Les enjeux d'une cartographie des SIC pour la discipline et les unités de recherche », *Revue française des sciences de l'information et de la communication* [En ligne], 5 | URL : http://journals.openedition.org/rfsic/1191

[5] Fiévet, P., & Guyot, F. (2018). *Automatic Categorization of Patent Documents in the International Patent Classification (IPCCAT)*. Présenté à The International Conference on Search, Data and Text Mining and Visualization. (IC-SDV), Nice. Consulté à l'adresse https://haxel.com/ii-sdv/2018/Programme/monday-23-april-2018

- *Clara Galliano : Laboratoire IMSIC – EA 7492*
  E-mail : clara-galliano@etud.univ-tln.fr
- *David Reymond : Laboratoire IMSIC – EA 7492*
  E-mail : dreymond@univ-tln.fr
- *Luc Quoniam : Laboratoire IMSIC – EA 7492*
  E-mail : mail@quoniam.info

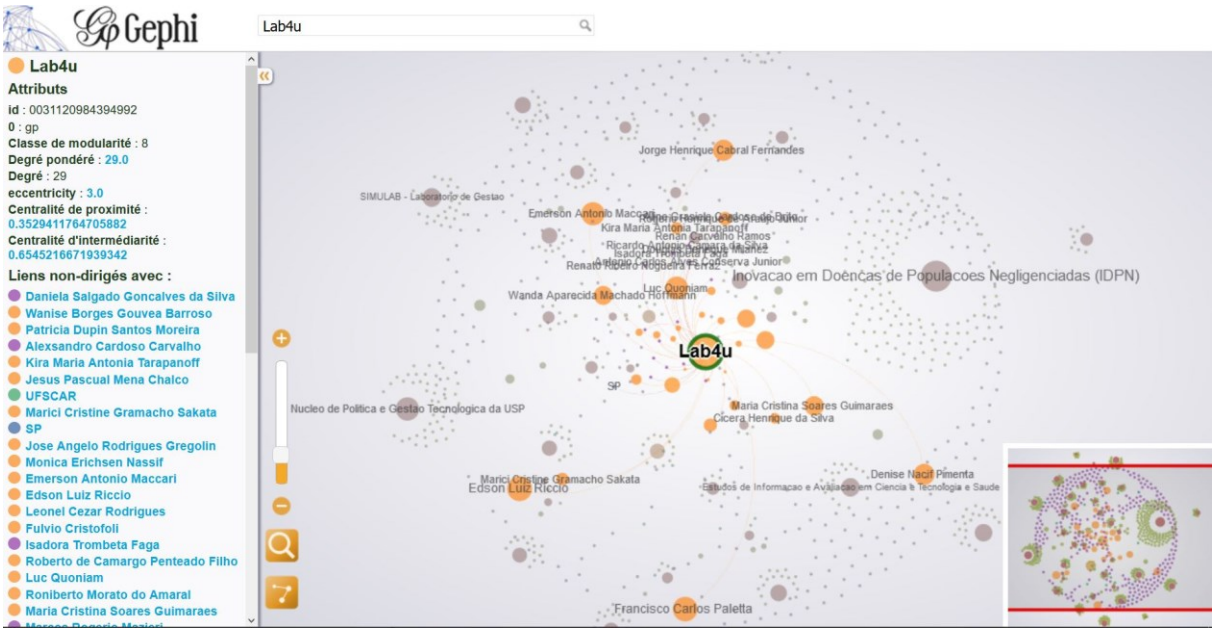

Figure 1 : exemple de réseau réalisé à partir du logiciel Gephi (recherche académique au Brésil à retrouver sur http://vlab4u.info/)

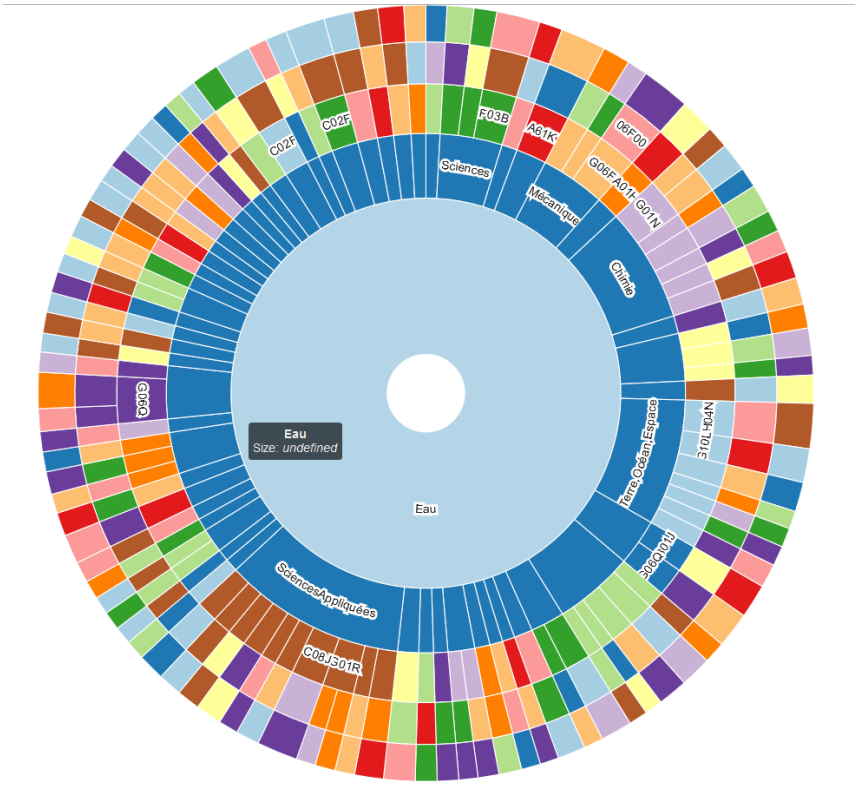

Figure 2 : exemple de visualisation proposé pour l'indexation des thèses françaises avec le mot clé "eau" via la classification internationale de la CIB (graphique Sunburst)